

Detection of Planted Solutions for Flat Satisfiability Problems

QUENTIN BERTHET^{*,‡} AND JORDAN S. ELLENBERG[†]

California Institute of Technology and University of Wisconsin

Abstract. We study the detection problem of finding planted solutions in random instances of flat satisfiability problems, a generalization of boolean satisfiability formulas. We describe the properties of random instances of flat satisfiability, as well of the optimal rates of detection of the associated hypothesis testing problem. We also study the performance of an algorithmically efficient testing procedure. We introduce a modification of our model, the light planting of solutions, and show that it is as hard as the problem of learning parity with noise. This hints strongly at the difficulty of detecting planted flat satisfiability for a wide class of tests.

Key words and phrases: Flat satisfiability, High-dimensional learning, Polynomial-time algorithms.

1. INTRODUCTION

The rapid growth in many scientific fields of the size of typical datasets, and the increasingly complex models that are studied, have naturally brought forth the notions of statistical and computational complexity in learning theory. For many learning problems motivated by such applications, the algorithmic aspect of inference procedures cannot be ignored: it is necessary to consider jointly the difficulties posed by the presence of noise or random errors, and by computational hardness.

The problem of understanding the tradeoffs between algorithmic and statistical efficiency, has therefore attracted a lot of interest. A particularly successful approach has been to investigate the links between learning problems that naturally arise, inspired by applications, and more abstract problems related to random discrete structures, that have been extensively studied in theoretical computer science. An hypothesis of [Fei02], based on the hardness of refuting satisfiability in random satisfiability formulas - initially used to prove hardness of approximation for several problems - has been used as a primitive to show hardness of improper learning [DLSS12, DLSS13, LSSS14]. An hypothesis on the planted clique problem has also been used as a primitive to prove computational limits to inference, initially for sparse principal component detection

^{*}Partially supported by AFOSR grant FA9550-14-1-0098.

[†]Partially supported by NSF grant DMS-1402620 and AFOSR Grant “Mathematical Foundations of Secure Computing Clouds”

[‡]The authors would like to thank Piyush Srivastava for very helpful discussions.

in [BR13], and subsequently for other problems in high dimensional statistics [MW13, Che13, WBS14, GMZ15, CLR15].

The desire to understand barriers to learning that come from randomness and computation has therefore brought attention to these problems, and the questions of learning distributions of their instances, in a computationally efficient manner. Examples include [FGR⁺13, FPV13, FK14, FPV14], investigating the query complexity of statistical algorithms for these problems [Kea98], or [Ber14] treating the problem of satisfiability detection as an hypothesis testing problem.

We consider here a learning problem on sets of flats in \mathbb{F}_2^n , shown to be a generalization of the k -SAT problem in n variables. We introduce the k -FLAT problem over sets of m flats of dimension $n - k$, that are flat satisfiable if they do not cover all of \mathbb{F}_2^n . This is analogous to satisfiability formulas, that are satisfiable if the m clauses do not exclude all the assignments. We also introduce a learning problem over these instances. It is formulated as a high-dimensional inference problem of hypothesis testing between the uniform distribution and the planted distribution, where an unknown assignment of \mathbb{F}_2^n is made flat-satisfiable. We study the optimal rates of detection for this problem, in a minimax sense, based on various parameters. We show that the optimal sample size m scales linearly with the dimension n . Even if they are derived considering only information-theoretic limits, these rates are useful as benchmarks. They give a context to the performance of candidate algorithms, and let us see if there is a gap between what we are able to achieve and the best possible case. We introduce a polynomial-time algorithm for a test, inspired by a technique of [AG11], and show that the test is successful for a sample of order n^k . We discuss how a modification of the problem, denoted by lightly planted flat satisfiability - that does not significantly alter it from a purely statistical point of view - affects the computational aspects, making it as hard as the “Learning Parity with Noise” problem [BKW03]. We also show how this result strongly suggests that a wide class of testing methods cannot be used for detection of planted solutions for flat satisfiability.

This article is structured in the following manner: Section 2 is focused on the description of these problems. We introduce the k -FLAT problem, and the associated problem of detecting planted solutions. In Section 3, we show that there exists a sharp phase transition for flat satisfiability of random instances, with a threshold at an explicit constant Δ in the linear regime $m = \Delta n$. In Section 4, we use this result to derive the optimal rate of detection, with an optimal constant, that coincides with the flat satisfiability transition. In Section 5, we show that a test that can be computed in polynomial time will be successful with a sample size that is polynomial in n . We introduce in Section 6 the problem of detecting a lightly planted solution, for which we describe optimal rates of detection, and discuss computational aspects.

2. PROBLEM DESCRIPTION

2.1 The k -FLAT problem

Consider \mathbb{F}_2^n , the n -dimensional coordinate space on \mathbb{F}_2 . We are given $V = (V_1, \dots, V_m)$, a collection of m flats of dimension $n - k$, or k -flats on \mathbb{F}_2^n . We denote by k -FLAT the problem of determining whether there exists an element $x \in \mathbb{F}_2^n$ that is *flat satisfying*, i.e. that does not lie on any of the V_j , or alternatively, whether $\mathbb{F}_2^n = \cup_j V_j$. We can define the flats by taking k linearly independent

linear forms $\ell_{j,1}, \dots, \ell_{j,k}$ and k values $\varepsilon_{j,1}, \dots, \varepsilon_{j,k} \in \mathbb{F}_2$, and having

$$V_j = \{x \in \mathbb{F}_2^n : \ell_{j,i}(x) = \varepsilon_{j,i}, \forall i \in [k]\}.$$

We note that there are many such descriptions for any flat, but choosing the $\ell_{j,i}$ and $\varepsilon_{j,i}$ uniformly at random does yield the uniform distribution on flats. We also note that if we constrain the flats to be coordinate-aligned by taking each linear form among the projections on one of the e_i s, the V_j can be interpreted as satisfiability clauses on k literals, and the set V_1, \dots, V_m a satisfiability formula with m clauses: For each $x \in \mathbb{F}_2^n$, x satisfies the j -th clause if and only if $x \notin V_j$, and satisfies the formula if and only if it the case for all the V_j . The set of flat satisfying assignments is therefore $\mathbb{F}_2^n \setminus \cup_j V_j$. The problem described above is therefore a generalization of k satisfiability. Thus, the k -FLAT problem is NP-complete for $k \geq 3$.

We denote by $\mathcal{S}(V)$ the set of flat satisfying elements $\mathbb{F}_2^n \setminus \cup_j V_j$, and by $Z(V)$ its cardinality. We write \mathcal{S} and Z when it is not ambiguous. We denote by FLAT the set of V that are flat satisfiable, i.e. for which there exists a satisfying element. We will consider asymptotics in the linear regime of $m = \Delta n$, for a constant $\Delta > 0$, and $m, n \rightarrow +\infty$.

2.2 Detection of planted flat-satisfiable assignment

Given a random instance V , our goal is to distinguish two hypotheses for its underlying joint distribution. This detection problem is a generalization of the problem of detecting planted satisfiability [Ber14]. Under the uniform distribution (denoted by \mathbf{P}_{unif}) the V_j s are independent and identically distributed. Their distribution is uniform on the set of flats of dimension $n-k$. A possible way to generate them is to draw uniformly k linearly independent linear forms $\ell_{j,1}, \dots, \ell_{j,k}$ and independently k values $\varepsilon_{j,1}, \dots, \varepsilon_{j,k} \in \mathbb{F}_2$, and to define

$$V_j = \{x \in \mathbb{F}_2^n : \ell_{j,i}(x) = \varepsilon_{j,i}, \forall i \in [k]\}.$$

Under the planted distribution, (denoted by $\mathbf{P}_{\text{planted}}$), an element $x^* \in \mathbb{F}_2^n$ is chosen uniformly. Conditioned on this element, the V_j s are independent and identically distributed, with a distribution denoted by \mathbf{P}_{x^*} . Under this distribution, they are chosen uniformly on the set of flats of dimension $n-k$ that do not contain x^* . They can be generated in a similar manner as under the uniform distribution, by drawing uniformly k linearly independent linear forms $\ell_{j,1}, \dots, \ell_{j,k}$, and the k values $\varepsilon_{j,i}$ uniformly among the $2^k - 1$ choices that are not all $\ell_{j,i}(x^*)$. We define V_j similarly. By construction, it does not contain x^* , which is a satisfying assignment for V .

REMARK 2.1. *Let G be the subgroup of $\mathbf{GL}_n(\mathbb{F}_2)$ consisting of linear transformations fixing x^* . Then G acts transitively on the k -flats not containing x^* . In particular, a probability distribution on k -flats which is supported on k -flats not containing x^* , and which is invariant under G , must be uniform on the k -flats not containing x^* ; in other words, it is the distribution \mathbf{P}_{x^*} described above. In particular, the procedure of choosing k linear forms ℓ_i and k bits ε_i uniformly at random subject to the conditions that the ℓ_i are linearly independent, and that the*

$\ell_i(x^*) - \varepsilon_i$ is nonzero for at least one i , is evidently G -invariant; thus, the resulting distribution on k -flats is \mathbf{P}_{x^*} . In this paper we will mostly use this description of \mathbf{P}_{x^*} . But we want to emphasize that there are many such descriptions, i.e. many distributions on k -tuples of pairs (ℓ, ε) which yield the distribution \mathbf{P}_{x^*} on k -flats. For instance, we could choose the ℓ_i as above, but then choose an i at random, require that $\ell_i(x^*) - \varepsilon_i = 1$, and allow the other $k-1$ bits ε_j to be chosen independently at random. Or we could require $\ell_i(x^*) - \varepsilon_i = 1$ for all i . Any of these processes result in a G -invariant distribution on k -flats not containing x^* , which can only be \mathbf{P}_{x^*} .

In order to avoid confusion regarding the representation of these flats, we consider here that the input data is the actual flat, given to us either as a membership oracle - a function that returns whether any element of \mathbb{F}_2^n belongs to the flat V_j - or as a uniformly random base ℓ_j of the space of linear forms that are constant on the flat, and the corresponding values ε_j . From a purely statistical point of view, this makes no difference: it is equivalent to consider a membership oracle, or the finite list of the elements of V_j , or the basis described here. From an algorithmic point of view, we will consider that our data is a uniformly random basis of linear forms and the associated values (ℓ_j, ε_j) for the k -flat, which has then the distribution described above.

Formally, we denote by q_0 the uniform distribution on k -flats of in \mathbb{F}_2^n , and for all $x \in \mathbb{F}_2^n$ by q_x the uniform distribution on k -flats of \mathbb{F}_2^n , that do not contain x . With these notations, the distributions considered in this problem are defined thus

$$\mathbf{P}_{\text{unif}} := q_0^{\otimes m}, \quad \mathbf{P}_{x, \pi} := q_x^{\otimes m}, \quad \mathbf{P}_{\text{planted}} := \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} \mathbf{P}_{x^*}.$$

Our detection problem can be written as testing between two hypotheses

$$\begin{aligned} H_0 &: V = (V_1, \dots, V_m) \sim \mathbf{P}_{\text{unif}} \\ H_0 &: V = (V_1, \dots, V_m) \sim \mathbf{P}_{\text{planted}}. \end{aligned}$$

3. FLAT-SATISFIABILITY THRESHOLD

In this section, we study the probability that a uniformly random instance V of the k -FLAT problem is flat satisfiable, when $m = \Delta n$, as a function of $\Delta > 0$. This is achieved by studying the first two moments of $Z(V)$.

LEMMA 3.1. *Under the uniform distribution*

$$\mathbf{E}[Z] = 2^n(1 - 2^{-k})^m.$$

PROOF. It holds that

$$Z = \sum_{x \in \mathbb{F}_2^n} \prod_{i=1}^m \mathbf{1}\{x \notin V_i\}.$$

By linearity, symmetry of the distribution, and independence of the V_j , for any $x_0 \in \mathbb{F}_2^n$

$$\mathbf{E}[Z] = 2^n(\mathbf{P}_{\text{unif}}(x_0 \notin V_1))^m.$$

Furthermore, for each k -flat of \mathbb{F}_2^n , $|V_1| = 2^{n-k}$, which yields the desired result. \square

LEMMA 3.2. *Let $V = (V_1, \dots, V_m)$ be a random collection of m k -flats on \mathbb{F}_2^n with distribution \mathbf{P}_{unif} . Let $m = \Delta n$, for some $\Delta > 0$. We have*

$$\frac{\mathbf{E}[Z^2]}{\mathbf{E}[Z]^2} \leq 1 + o(1) + \frac{1}{\mathbf{E}[Z]}.$$

PROOF. We derive the second moment of Z

$$\begin{aligned} Z^2 &= \sum_{x, x' \in \mathbb{F}_2^n} \mathbf{1}\{x \in \mathcal{S}(V)\} \mathbf{1}\{x' \in \mathcal{S}(V)\} \\ &= \sum_x \mathbf{1}\{x \in \mathcal{S}(V)\} + \sum_{x \neq x'} \mathbf{1}\{x \in \mathcal{S}(V)\} \mathbf{1}\{x' \in \mathcal{S}(V)\}. \end{aligned}$$

Taking expectation yields

$$\mathbf{E}[Z^2] = \mathbf{E}[Z] + \sum_{x \neq x'} \mathbf{P}_{\text{unif}}(\{x \in \mathcal{S}(V)\} \cap \{x' \in \mathcal{S}(V)\}).$$

The uniform distribution is invariant under the action of $GL_n(\mathbb{F}_2)$, which is doubly transitive on \mathbb{F}_2^n . Therefore, the term $\mathbf{P}_{\text{unif}}(\{x \in \mathcal{S}(V)\} \cap \{x' \in \mathcal{S}(V)\})$ is constant for all couples of distinct elements (x, x') of \mathbb{F}_2^n . To compute this distribution, it thus suffices to consider that x and x' are uniformly randomly chosen among the set of pairs of distinct elements. For all $j \in [m]$, this yields

$$\mathbf{P}_{\text{unif}}(\{x \notin V_j\} \cap \{x' \notin V_j\}) = \frac{2^n - 2^{n-k}}{2^n} \cdot \frac{2^n - (2^{n-k} - 1)}{2^n - 1} = (1 - 2^{-k}) \left(1 - 2^{-k} + \frac{2 + 2^{-k}}{2^n - 1}\right).$$

Using this in the derivation of the second moment, we have

$$\begin{aligned} \mathbf{E}[Z^2] &= \mathbf{E}[Z] + (2^{2n} - 2^n)(1 - 2^{-k})^m \left(1 - 2^{-k} + \frac{2 + 2^{-k}}{2^n - 1}\right)^m \\ &\leq \mathbf{E}[Z] + 2^{2n}(1 - 2^{-k})^{2m} \left(1 + \frac{2 + 2^{-k}}{1 - 2^{-k}} \frac{1}{2^n - 1}\right)^m \\ &\leq \mathbf{E}[Z] + \mathbf{E}[Z]^2 \left(1 + \frac{2 + 2^{-k}}{1 - 2^{-k}} \frac{1}{2^n - 1}\right)^{\Delta n}. \end{aligned}$$

Note that the last term is a $1 + o(1)$. □

Together, Lemma 3.1 and 3.2 yield the following

THEOREM 3.3. *For $k > 0$ let $\Delta_k := \log(1/2)/\log(1 - 2^{-k}) \approx 2^k \log(2)$. For $\Delta > 0$, let $m = \Delta n$, and V be uniformly distributed. When $m, n \rightarrow +\infty$, it holds that*

- For $\Delta < \Delta_k$, $\mathbf{P}_{\text{unif}}(V \in \text{FLAT}) \rightarrow 1$.
- For $\Delta > \Delta_k$, $\mathbf{P}_{\text{unif}}(V \in \text{FLAT}) \rightarrow 0$.

PROOF. We first note that $2(1 - 2^{-k})^{\Delta_k} = 1$, so that $\mathbf{E}[Z] = [2(1 - 2^{-k})^\Delta]^n$ is exponentially large when $\Delta < \Delta_k$, and exponentially small when $\Delta > \Delta_k$.

- For $\Delta < \Delta_k$, Markov's inequality yields

$$\mathbf{P}_{\text{unif}}(V \in \text{FLAT}) = \mathbf{P}_{\text{unif}}(Z(V) \geq 1) \leq \mathbf{E}[Z] \rightarrow 0.$$

- For $\Delta < \Delta_k$, Paley-Zigmund's inequality and the result of Lemma 3.2 yields

$$\mathbf{P}_{\text{unif}}(V \in \text{FLAT}) = \mathbf{P}_{\text{unif}}(Z(V) > 0) \geq \frac{\mathbf{E}[Z]^2}{\mathbf{E}[Z^2]} \rightarrow 1.$$

□

There is therefore a sharp phase transition in the linear regime, at Δ_k , where the limit of the probability of flat satisfiability switches from 1 to 0. This result can be compared to the satisfiability transition for k -SAT problems, for which Z has the same expectation, but for which the second moment is much larger than $\mathbf{E}[Z]^2$. The proofs of satisfiability transitions [AP04, COP13, DSS14] are therefore much more technical.

4. OPTIMAL DETECTION FOR PLANTED FLAT-SATISFIABILITY

One can understand the two distributions by the following generating process. Let \mathcal{N}_k be the number of subspaces of dimension $n - k$ in \mathbb{F}_2^n . There are therefore $2^k \mathcal{N}_k$ possible k -flats (equivalent to a choice of linear forms, and k values). Under the uniform distribution, m flats are chosen independently and uniformly among the $2^k \mathcal{N}_k$ possible choices. Under \mathbf{P}_{x^*} , there is an excluded choice of values, and there are $(2^k - 1)\mathcal{N}_k$ allowed flats, among which we draw independently and uniformly m flats. This interpretation of the distributions is useful to derive the likelihood ratio, in the following.

LEMMA 4.1. *Let $V = (V_1, \dots, V_m)$ be a collection of m k -flats on \mathbb{F}_2^n ,*

$$\frac{\mathbf{P}_{\text{planted}}(V)}{\mathbf{P}_{\text{unif}}(V)} = \frac{Z(V)}{\mathbf{E}[Z]}.$$

PROOF. By definition of $\mathbf{P}_{\text{planted}}$

$$\frac{\mathbf{P}_{\text{planted}}(V)}{\mathbf{P}_{\text{unif}}(V)} = \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} \frac{\mathbf{P}_{x^*}(V)}{\mathbf{P}_{\text{unif}}(V)}.$$

To compute the probabilities in the above ratios, we use the interpretation above of m drawings in $N = 2^k \mathcal{N}_k$ possible flats independently if the distribution is \mathbf{P}_{unif} , or otherwise in $N^* = (2^k - 1)\mathcal{N}_k$ possible choices corresponding to flats that do not contain x^* . Therefore, it holds for all V

$$\frac{\mathbf{P}_{x^*}(V)}{\mathbf{P}_{\text{unif}}(V)} = \begin{cases} 0 & \text{if } x \notin \mathcal{S}(V) \\ \left(\frac{N}{N^*}\right)^m & \text{otherwise} \end{cases}$$

Therefore, the likelihood ratio can be expressed in terms of $\mathbf{1}\{x \in \mathcal{S}(V)\}$, and $N/N^* = 1/(1 - 2^{-k})$

$$\begin{aligned} \frac{\mathbf{P}_{\text{planted}}}{\mathbf{P}_{\text{unif}}}(V) &= \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} \left(\frac{N}{N^*}\right)^m \mathbf{1}\{x \in \mathcal{S}(V)\} \\ &= \frac{1}{\mathbf{E}[Z]} \sum_{x \in \mathbb{F}_2^n} \mathbf{1}\{x \in \mathcal{S}(V)\} = \frac{Z(V)}{\mathbf{E}[Z]}. \end{aligned}$$

□

The distribution $\mathbf{P}_{\text{planted}}$ therefore has a likelihood proportional to $Z(V)$: only the flat satisfiable V have a positive measure, and those with a large number of flat satisfying assignments are more likely to occur. This can be contrasted with the uniform distribution on **FLAT**, for which all flat satisfiable V are equally likely. One of the motivations behind the study of this likelihood ratio is its relationship with the total variation distance. Indeed, we have

$$d_{\text{TV}}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted}}) = \frac{1}{2} \mathbf{E} \left[\left| \frac{Z}{\mathbf{E}[Z]} - 1 \right| \right] \leq \frac{1}{2} \sqrt{\frac{\mathbf{E}[Z^2]}{\mathbf{E}[Z]^2} - 1}.$$

The last inequality is a consequence of Jensen's inequality, and gives a more tractable bound on the total variation distance. It is equivalent to considering the χ^2 divergence between the two distributions. When $\Delta < \Delta_k$, Lemma 3.2 yields

$$d_{\text{TV}}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted}}) \leq \frac{1}{2} \sqrt{\frac{\mathbf{E}[Z^2]}{\mathbf{E}[Z]^2} - 1} \leq \frac{1}{2} \sqrt{\frac{1}{\mathbf{E}[Z]} + o(1)} \rightarrow 0.$$

Note that this approach is not fruitful to control the total variation distance in the k -SAT planted satisfiability problem, as $\mathbf{E}[Z^2]$ is too large, in the linear regime of $m = \Delta n$ for some constant $\Delta > 0$.

For this problem, when $\Delta > \Delta_k$, $\mathbf{P}_{\text{unif}}(Z > 0) \leq \mathbf{E}[Z] \rightarrow 0$. Checking flat satisfiability, i.e. if $Z > 0$ is therefore a test with a one-sided probability of error equal to $\mathbf{P}_{\text{unif}}(Z > 0)$, as we have $\mathbf{P}_{\text{planted}}(Z > 0) = 1$. Together, these two observations yield the following

THEOREM 4.2. *For a fixed $\Delta > 0$, let $m = \Delta n$. The following holds*

- For $\Delta > \Delta_k$, and $\psi_{\text{FLAT}}(V) = \mathbf{1}\{Z(V) > 0\}$

$$\mathbf{P}_{\text{unif}}(\psi_{\text{FLAT}} = 1) \vee \mathbf{P}_{\text{planted}}(\psi_{\text{FLAT}} = 0) \rightarrow 0.$$

- For $\Delta < \Delta_k$,

$$\inf_{\psi} \mathbf{P}_{\text{unif}}(\psi = 1) \vee \mathbf{P}_{\text{planted}}(\psi = 0) \rightarrow \frac{1}{2}.$$

We observe in the statistical problem the same phase transition as in Theorem 3.3: the problem switches at Δ_k from being insolvable (with a total variation distance converging to 0) to the existence of an powerful test, i.e. checking flat satisfiability. Note that in this regime, since $\mathbf{E}[Z] < 1$, this test is equivalent to the likelihood ratio test $Z(V) > \mathbf{E}[Z]$.

The picture is clear from the statistical and probabilistic point of view. However, from a computational point of view, checking if Z is equal to 0 (i.e. if the union of flats covers \mathbb{F}_2^n) is an **NP**-complete problem for $k \geq 3$, as k -SAT is a particular case. An interesting question is whether there are detection methods that can solve this problem in an algorithmically efficient manner.

5. POLYNOMIAL-TIME DETECTION

We study here the statistical performance of a test that runs in polynomial time. We introduce some notations necessary to define this test. Let W be a k -flat of \mathbb{F}_2^n , defined by k affine constraints

$$W = \{x \in \mathbb{F}_2^n : \ell_i(x) = \varepsilon_i, \forall i \in [k]\}.$$

We make the observation that x does not lie on W if and only if one of the above equations is not satisfied, or equivalently, taking $\alpha_i = 1 - \varepsilon_i$

$$x \notin W \iff P_{\ell, \alpha}(x) := \prod_{i=1}^k (\ell_i(x) + \alpha_i) = 0.$$

Factoring out, $P_{\ell, \alpha}$ can be written as a multivariate polynomial over \mathbb{F}_2 of degree k

$$P_{\ell, \alpha}(x) = \sum_{\substack{S \subset [n] \\ |S| \leq k}} c_S(\ell, \alpha) \prod_{s \in S} x_s.$$

Note that all the monomials are squarefree, as $z^2 = z$ for all $z \in \mathbb{F}_2$. Solving the k -FLAT problem is therefore equivalent to solving a system of m polynomial equations of degree k . Of course, this is an NP-hard problem. In order to obtain a test that is computationally tractable, we lift this system of equations in a higher dimensional space to obtain a system of linear equations with quadratic constraints, that we will then relax. This general idea is common over reals [Par01, Las01], and adapted here in a finite field. In this particular context, this approach is inspired by [AG11], where this technique is used in a problem of learning with errors.

Let $N_k = \sum_{i=0}^k \binom{n}{i} \leq (n+1)^k$, and for $x \in \mathbb{F}_2^n$, let $X \in \mathbb{F}_2^{N_k}$ such that $X_S = \prod_{s \in S} x_s$. We remark that $P_{\ell, \alpha}$ takes the same values as a linear form $\mathcal{L}_{\ell, \alpha}$ over $\mathbb{F}_2^{N_k}$, such that $P_{\ell, \alpha}(x) = \mathcal{L}_{\ell, \alpha}(X)$ for the X associated to x , by taking

$$\mathcal{L}_{\ell, \alpha}(X) = \sum_{\substack{S \subset [n] \\ |S| \leq k}} c_S(\ell, \alpha) X_S.$$

If we consider the mapping ϕ from \mathbb{F}_2^n to $\mathbb{F}_2^{N_k}$, the so-called Veronese embedding, that associates x to X , and $\mathcal{V} \subset \mathbb{F}_2^{N_k}$ the image of ϕ , it is equivalent to solve $P_{\ell, \alpha}(x) = 0$ over all of \mathbb{F}_2^n and $\mathcal{L}_{\ell, \alpha}(X) = 0$ over \mathcal{V} . In particular, determining if an instance of the k -FLAT problem is flat satisfiable is equivalent to determining if a system of m linear equations in $\mathbb{F}_2^{N_k}$ has a solution in \mathcal{V} . The image \mathcal{V} can be written as the intersection of quadratic constraints of the type $X_{\{1\}} X_{\{2\}} = X_{\{1,2\}}$, making the system of equations intractable. In order to obtain a tractable approximation of this problem, we consider the relaxed linear system of equations, by keeping solely the constraint $X_\emptyset = 1$. Formally, for an instance V of the k -FLAT problem, we will consider for each flat V_j the associated linear form $\mathcal{L}_{\ell_j, \alpha_j}$, and the overall system \mathcal{L}_V of $m+1$ linear equations in $\mathbb{F}_2^{N_k}$

$$(\mathcal{L}_V) \quad \mathcal{L}_{\ell_j, \alpha_j}(X) = 0, \forall j \in [m]; \quad X_\emptyset = 1.$$

Note that if $x^* \in \mathbb{F}_2^n$ is flat-satisfiable for V , the associated $X^* = \phi(x^*) \in \mathbb{F}_2^{N_k}$ is a solution to \mathcal{L}_V , as it is even a solution to the linear system of equations with stricter constraint $X \in \mathcal{V}$. As a consequence, the system \mathcal{L}_V always has a solution for $V \sim \mathbf{P}_{\text{planted}}$. However, under the uniform distribution, it is not always the case.

LEMMA 5.1. *Recall that $\Delta_k := \log(1/2)/\log(1 - 2^{-k}) \approx 2^k \log(2)$. Let $m = \Delta N_k$ for $\Delta > \Delta_k$, and $V = (V_1, \dots, V_m) \sim \mathbf{P}_{\text{unif}}$. The linear system \mathcal{L}_V has no solutions in $\mathbb{F}_2^{N_k}$, with probability converging to 1 when $n \rightarrow +\infty$.*

PROOF. Consider a fixed $Z \in \mathbb{F}_2^{N_k}$ such that $Z_\emptyset = 1$. For an k -flat W described by (ℓ, α) , we write $\mathcal{L}_{\alpha, \ell}(Z)$ as a function $q_{Z, \ell}$ of $\alpha \in \mathbb{F}_2^k$

$$q_{Z, \ell}(\alpha) = \sum_{\substack{S \subseteq [n] \\ |S| \leq k}} c_S(\ell, \alpha) Z_S.$$

We observe that each $c_S(\ell, \cdot)$ is a multivariate multilinear polynomial (with monomials that are squarefree), so that $q_{Z, \ell} = \mathbb{F}_2[\alpha_1, \dots, \alpha_k]$. Furthermore, the coefficient of the monomial $\alpha_1 \dots \alpha_k$ is $Z_\emptyset = 1$. As the squarefree monomials are linearly independent, there exists an element of \mathbb{F}_2^k such that $q_{Z, \ell}(\alpha) \neq 0$. Therefore, as α is uniformly distributed under the uniform distribution q_0 , it holds that

$$\mathbf{P}_{\text{unif}}(\mathcal{L}_{\alpha, \ell}(Z) = 0) = \mathbf{P}_{\text{unif}}(q_{Z, \ell}(\alpha) = 0) \leq 1 - 2^{-k}.$$

As an aside, note that this bound is tight. Indeed, for all $Z \in \mathcal{V}$, the event $\mathcal{L}_{\alpha, \ell}(Z) = 0$ is equivalent to $z \notin W$, for $z = \phi^{-1}(Z)$. The probability of this event is $1 - 2^{-k}$, as seen in the proof of Lemma 3.1.

Let $V = (V_1, \dots, V_m) \sim \mathbf{P}_{\text{unif}}$. By independence, we obtain directly that

$$\mathbf{P}_{\text{unif}}(\mathcal{L}_{\ell_j, \alpha_j}(X) = 0, \forall j \in [m]) \leq (1 - 2^{-k})^m.$$

By a union bound over all elements of $\mathbb{F}_2^{N_k}$, it holds that

$$\mathbf{P}_{\text{unif}}(\mathcal{L}_V \text{ has a solution}) \leq 2^{N_k} (1 - 2^{-k})^m.$$

Taking $\Delta > \Delta_k$ yields the desired result. \square

We consider the test $\psi_{\mathcal{L}} : V \mapsto \mathbf{1}\{\mathcal{L}_V \text{ has a solution}\}$. When m is of order $N_k \leq (n+1)^k$, it is possible to construct and solve the linear system, and thus to determine the outcome of the test, in time $O(n^{3k})$, by Gaussian elimination. The result of Lemma 5.1 gives a guarantee, in terms of sample size, about the performance of this test.

THEOREM 5.2. *Let $m = \Delta n^k$, for $\Delta > \Delta_k$. It holds that*

$$\mathbf{P}_{\text{unif}}(\psi_{\mathcal{L}} = 1) \vee \mathbf{P}_{\text{planted}}(\psi_{\mathcal{L}} = 0) \rightarrow 0.$$

There are several remarks that one can make about this result. The test $\psi_{\mathcal{L}}$ allows to distinguish the two distributions with probability of error going to 0, with computation time and sample size that are both polynomial in n . In particular, we show that the sample size m needs only to be of order n^k , which can be compared to results in [AG11], where this linearization procedure is shown to recover an analogue to the planted assignment x^* , with sample size n^{2k} . The statistical performance shown here is however suboptimal, and it is not clear whether there exists a test that runs in time polynomial in n and that can distinguish the two distributions with high probability for a sample size linear in n , the optimal regime, that can be seen as a benchmark.

There are other detection problems for which the optimal regime of detection is not known to be attainable by algorithmically efficient testing methods. In particular, for the planted clique problem [Jer92, Kuč95] in a graph with n vertices, even though cliques of size greater than $2\log_2(n)$ can be detected or recovered, polynomial-time algorithms are only known to be efficient at size \sqrt{n} [AKS98], widely believed to be optimal. This hypothesis has recently been used as a primitive to show hardness for other learning problems. This problem, as well as those of estimating planted assignments for CSP problems have been studied, and computational lower bounds shown to exist, in a specific computational model [FGR⁺13, FPV13].

A common type of method to solve these detection problems, one that comes naturally to mind to find an improved algorithm for this problem - i.e. that would need significantly less than n^k samples - is to study the behavior of a judiciously chosen, tractable statistic Σ of the data D . When D is constituted of m independent samples, let us consider only Σ that are sums of statistics ρ of r -tuples of the data, for a finite r . Simply, these approaches revolve around showing that $\Sigma(D)$ behaves differently under the two distributions of interest, say $\mathbf{E}_{\text{uniform}}[\Sigma(D)] = 0$, and $\mathbf{E}_{\text{planted}}[\Sigma(D)] = \mu > 0$, and by showing that when the sample size is large enough, μ is much greater than the typical deviations of Σ , making a test such as $\mathbf{1}\{\Sigma(D) > \mu/2\}$ powerful. Typical examples include statistics based on the degrees of vertices in a graph, bias in signs of literals in a CSP, etc.

This is not the approach used here, where our test is based on the *existence* of an element verifying certain properties - here being a solution to a linear system of equations in a finite field - not on summing a certain statistic over i.i.d samples (or couples, or triplets of these samples). In the following section, we describe a modified version of our hypothesis testing problem, by introducing the model of light planting. We show that even though it does not change the statistical nature of the problem, it is as hard as the “Learning Parity with Noise” problem, strongly suggesting that it cannot be efficiently solved. Therefore, it is highly improbable that any method that is robust to this modification - which is often true for the approaches based on biases of statistics, as described above - could be successful for detection of planted flat satisfiability.

6. DETECTION OF LIGHTLY PLANTED FLAT-SATISFIABILITY

We consider a modified version of our hypothesis testing problem. It has the same null hypothesis and in the alternative, planting only happens with some constant probability $\pi \in (0, 1)$, which we call light planting. Formally, we denote

by $q_{x,\pi} := (1 - \pi)q_0 + \pi q_x$ the distribution on the flats of dimension $n - k$ that is mixture of the uniform q_0 and of the planting distribution q_x , and define similarly $\mathbf{P}_{x,\pi}$ and $\mathbf{P}_{\text{planted},\pi}$. As in the planting model with $\pi = 1$, we have

$$\mathbf{P}_{x,\pi} := q_{x,\pi}^{\otimes m}, \quad \mathbf{P}_{\text{planted},\pi} := \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} \mathbf{P}_{x,\pi}.$$

The alternative hypothesis is therefore replaced with $H_{1,\pi} : V = (V_1, \dots, V_m) \sim \mathbf{P}_{\text{planted},\pi}$.

6.1 Optimal rates of detection for light planting

To tackle this problem, we consider for a given set of flats V the following statistics

$$s(V, x) = |\{j : x \notin V_j\}|, \text{ and } \sigma(V) = \max_{x \in \mathbb{F}_2^n} s(V, x).$$

They are respectively the number of flats of V on which x does not lie, and the maximum number of flat constraints simultaneously satisfiable by an element of \mathbb{F}_2^n . We derive the following deviation bounds for this second statistic under both hypotheses.

LEMMA 6.1. *For a fixed $\Delta > 0$, let $m = \Delta n$. It holds that*

$$\begin{aligned} \mathbf{P}_{\text{unif}}(\sigma(V) > [(1 - 2^{-k}) + \alpha]m) &\leq \exp(-[2\alpha^2\Delta - \log(2)]n) \\ \mathbf{P}_{\text{planted},\pi}(\sigma(V) < [(1 - 2^{-k}) + \pi 2^{-k} - \alpha]m) &\leq \exp(-2\alpha^2\Delta n). \end{aligned}$$

PROOF. For all $x \in \mathbb{F}_2^n$, we observe that under the null hypothesis, the variable $s(x, V)$ has distribution $\mathcal{B}(m, 1 - 2^{-k})$. Therefore, by Hoeffding's inequality,

$$\mathbf{P}_{\text{unif}}(s(x, V) > [(1 - 2^{-k}) + \alpha]m) \leq \exp(-2\alpha^2m).$$

A union bound on \mathbb{F}_2^n yields

$$\mathbf{P}_{\text{unif}}(\sigma(V) > [(1 - 2^{-k}) + \alpha]m) \leq 2^n \exp(-2\alpha^2m) \leq \exp(-[2\alpha^2\Delta - \log(2)]n).$$

Under $\mathbf{P}_{x^*,\pi}$ the variable $s(x^*, V)$ has distribution $\mathcal{B}(m, (1 - 2^{-k}) + \pi 2^{-k})$. By Hoeffding's inequality,

$$\mathbf{P}_{x^*,\pi}(s(x^*, V) < [(1 - 2^{-k}) + \pi 2^{-k} - \alpha]m) \leq \exp(-2\alpha^2m).$$

By definition of $\mathbf{P}_{\text{planted},\pi}$ and $\sigma(V) \geq s(x, V)$ for all $x \in \mathbb{F}_2^n$, we obtain the desired result. \square

These deviation can be used to prove that a particular test is powerful in the linear regime.

THEOREM 6.2. *For a fixed $\Delta > 0$, let $m = \Delta n$, $\tilde{\Delta}_{k,\pi} := 2^{2k-1} \log(2)/\pi^2$ and $\Delta_{k,\pi} := 2^k \log(2)/\pi^2$. It holds that*

- For $\Delta > \tilde{\Delta}_{k,\pi}$, and $\psi_\sigma(V) = \mathbf{1}\{\sigma(V) > [(1 - 2^{-k}) + \pi 2^{-(k+1)}]m\}$

$$\mathbf{P}_{\text{unif}}(\psi_\sigma = 1) \vee \mathbf{P}_{\text{planted},\pi}(\psi_\sigma = 0) \rightarrow 0.$$

- For $\Delta < \Delta_{k,\pi}$,

$$\inf_{\psi} \mathbf{P}_{\text{unif}}(\psi = 1) \vee \mathbf{P}_{\text{planted},\pi}(\psi = 0) \rightarrow \frac{1}{2}.$$

The first point of this result is a direct consequence of Lemma 6.1. The proof of the second point is based on a bound between the χ^2 divergence of the two distributions, similarly to the result in Theorem 4.2. A full proof of the theorem can be found in Appendix A. If we consider π to be a constant, the optimal rate of detection for the light planting version of the problem is therefore still in the linear regime $m = \Delta_{k,\pi}n$. Furthermore the right dependency of $\Delta_{k,\pi}$ on π is in $1/\pi^2$, up to constants that only depend on k .

6.2 Computational aspects of light planting

The algorithmically efficient testing method described in Section 5 is not robust to this modification of the hypothesis testing problem: it relies heavily on the fact that for $V \sim \mathbf{P}_{\text{planted}}$, there exists some x^* that is flat-satisfiable, which guarantees in turn the existence of a solution to the linear system \mathcal{L}_V . This reasoning does not go through under the light planting model.

We give here strong reasons to believe that improving the result of Theorem 5.2 - for the case $\pi = 1$ - by using this type of method is hopeless. Our reasoning is that such an approach would be robust to light planting, and would allow to distinguish \mathbf{P}_{unif} and $\mathbf{P}_{\text{planted},\pi}$ with sample size and running time polynomial in n . The following result shows that this would imply in turn the existence of an efficient method for the decision version of the “Learning Parity with Noise” (LPN) problem of [BKW03], known to be as hard as the recovery of the “secret” signal. This is conjectured to be a hard problem, for which the best algorithms run in time $2^{O(n/\log(n))}$, and used to prove the safety of cryptography systems (see [Pie12], and references within).

Let $(A, b) \in \mathbb{F}_2^{n \times m} \times \mathbb{F}_2^m$ be an instance of LPN. For each $j \in [m]$, let $\gamma_{j,1}, \dots, \gamma_{j,k-1}$ be $k-1$ uniformly random, linearly independent linear forms of \mathbb{F}_2^n , themselves independent of the linear form φ_j generated by A_j . If A_j is uniformly random, the $n-k$ dimensional linear subspace of \mathbb{F}_2^n that is the vanishing set of these k linear forms is therefore uniformly random as well. Furthermore, let $\beta_{j,1}, \dots, \beta_{j,k-1}$ be $k-1$ independent, uniformly random elements of \mathbb{F}_2 , independent of b_j . Take $\ell_{j,1}, \dots, \ell_{j,k}$ be equal to $\gamma_{j,1}, \dots, \gamma_{j,k-1}, \varphi_j$ in a uniformly random order, and $\varepsilon_{j,1}, \dots, \varepsilon_{j,k}$ be equal to $\beta_{j,1}, \dots, \beta_{j,k-1}, 1 - b_j$ in the same order. The equation $\ell_j(x) = \varepsilon_j$ defines the $n-k$ dimensional flat V_j .

LEMMA 6.3. *Let (A, b) be an instance of LPN, and V the associated instance of k -FLAT obtained by the procedure described above. The following holds*

- If (A, b) are independent and uniformly random, $V \sim \mathbf{P}_{\text{unif}}$.
- If (A, b) is an instance with secret x , and probability of error $\eta < 1/2$, $V \sim \mathbf{P}_{x,\pi}$, with $\pi = 1 - 2\eta$.

PROOF. In all cases, the k -flats are independent, and the m sets of k linear forms are uniformly distributed. If (A, b) is uniformly random, so are the b_j , and as a consequence, the ε_j . This yields the desired $V \sim \mathbf{P}_{\text{unif}}$. However, if there is a secret x , $\phi_j(x) = 1 - b_j$ with probability η . The distribution of $1 - b_j -$

$\phi_j(x)$ is therefore is a mixture of the uniform distribution on \mathbb{F}_2 (with weight $1 - \pi$) and of the unit mass at 1 (with weight π). The distribution of $\varepsilon_j - \ell_j(x)$ is thus the mixture of the uniform distribution on \mathbb{F}_2^n (with weight $1 - \pi$) and of the the distribution on $\mathbb{F}_2^k \setminus \{0\}$ generated by placing a 1 in one of the coefficients of $\varepsilon_j - \ell_j(x)$, and letting the others be independent and uniform. As shown in Remark 2.1, the resulting flat V_j has distribution $q_{x,\pi}$ and $V \sim \mathbf{P}_{x,\pi}$, as desired. \square

From a computational point of view, there is a very strong difference between the problems of detecting planted solutions to flat satisfiability, and detecting solutions that are only lightly planted, for any constant $\pi \in (0, 1)$. It seems impossible to adapt the result of Theorem 5.2 to this new setting, and to describe an efficient algorithm that can distinguish these distributions for a sample size of order n^k/π^2 , similarly to the result of Theorem 6.2.

The testing methods based on simple statistics (i.e. sums of simpler statistics that depend on finite r -tuples of samples) as described in Section 5, are usually robust to these modifications. As an example, for the planted clique problem, consider a light planting distribution that only plants edges in the small subgraph with probability π . The sum of the degrees of all the vertices has mean $\frac{n(n-1)}{4}$ under the null, and respectively $\frac{n(n-1)}{4} + \frac{k(k-1)}{2}$ and $\frac{n(n-1)}{4} + \pi \frac{k(k-1)}{2}$ under the planted, or lightly planted distribution. Deviation bounds will therefore show that a test based on this statistic will be successful when $k \geq C\sqrt{n}$ under the planted model and $k \geq C\sqrt{n/\pi}$ under the lightly planted model, for some constant $C > 0$. The rates of detection for this method are not changed by this modification, for a constant π . The situation is similar for detection of planted satisfiability [Ber14, Thm 3.1]: a statistic based on joint signs of variables appearing several times in the formula has mean 0 under the uniform distribution, and mean $1/[2(2^k - 1)]$ under the planted distribution, and would have mean $\pi^2/[2(2^k - 1)]$ under the light planting model. The necessary sample size m of order \sqrt{n} in this problem would only be affected in the constant by π .

Here, this problem is significantly harder to solve in an algorithmically efficient manner when light planting is introduced. Any candidate algorithm to solve the planting problem (with $\pi = 1$) would need therefore not be of the type informally described above, and need to not be robust to this type of modification in the distributions. Indeed, an algorithm robust to light planting that is statistically and algorithmically efficient could otherwise be used to solve the LPN problem, as shown in Lemma 6.3.

REFERENCES

- [AG11] Sanjeev Arora and Rong Ge, *New algorithms for learning in the presence of errors*, ICALP (2011).
- [AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Proceedings of the Eighth International Conference “Random Structures and Algorithms” (Poznan, 1997), vol. 13, 1998, pp. 457–466. [MR1662795 \(99k:05144\)](#)
- [AP04] Dimitris Achlioptas and Yuval Peres, *The threshold for random k -sat is $2^k \ln 2 - o(k)$* , J. Amer. Math. Soc. **17** (2004), 947–973.

- [Ber14] Quentin Berthet, *Optimal testing for planted satisfiability problems*, Electron. J. Stat., to appear (2014).
- [BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman, *Noise-tolerant learning, the parity problem, and the statistical query model*, J. ACM **50** (2003), no. 4, 506–519.
- [BR13] Quentin Berthet and Philippe Rigollet, *Complexity theoretic lower bounds for sparse principal component detection*, J. Mach. Learn. Res. (COLT) **30** (2013), 1046–1066.
- [Che13] Yudong Chen, *Incoherence-optimal matrix completion*.
- [CLR15] T. Tony Cai, Tengyuan Liang, and Alexander Rakhlin, *Computational and statistical boundaries for submatrix localization in a large noisy matrix*.
- [COP13] Amin Coja-Oghlan and Konstantinos Panagiotou, *Going after the k -sat threshold*, STOC '13 Proceedings of the 45th annual ACM symposium on Symposium on theory of computing (2013), 705–714.
- [DLSS12] Amit Daniely, Nati Linial, and Shai Shalev-Shwartz, *The complexity of learning halfspaces using generalized linear methods*.
- [DLSS13] ———, *From average case complexity to improper learning complexity*.
- [DSS14] Jian Ding, Allan Sly, and Nike Sun, *Proof of the satisfiability conjecture for large k* .
- [Fei02] Uriel Feige, *Relations between average case complexity and approximation complexity*, Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing (New York), ACM, 2002, pp. 534–543 (electronic). [MR2121179](#)
- [FGR⁺13] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao, *Statistical algorithms and a lower bound for planted clique*, Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, STOC 2013, 2013.
- [FK14] Vitaly Feldman and Pravesh Kothari, *Agnostic learning of disjunctions on symmetric distributions*.
- [FPV13] Vitaly Feldman, Will Perkins, and Santosh Vempala, *On the complexity of random satisfiability problems with planted solutions*, Arxiv Preprint (2013).
- [FPV14] ———, *Subsampled power iteration: a unified algorithm for block models and planted csp's*.
- [GMZ15] Chao Gao, Zongming Ma, and Harrison H. Zhou, *Sparse cca: Adaptive estimation and computational barriers*.
- [Jer92] Mark Jerrum, *Large cliques elude the Metropolis process*, Random Structures Algorithms **3** (1992), no. 4, 347–359. [MR1179827 \(94b:05171\)](#)
- [Kea98] Michael Kearns, *Efficient noise-tolerant learning from statistical queries*, J. ACM **45** (1998), no. 6, 983–1006.
- [Kuĉ95] Ludĉk Kuĉera, *Expected complexity of graph partitioning problems*, Discrete Appl. Math. **57** (1995), no. 2-3, 193–212, Combinatorial optimization 1992 (CO92) (Oxford). [MR1327775 \(96c:68091\)](#)
- [Las01] Jean B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM JOURNAL ON OPTIMIZATION **11** (2001),

- 796–817.
- [LSSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir, *On the computational efficiency of training neural networks*.
 - [MW13] Zongming Ma and Yihong Wu, *Computational barriers in minimax submatrix detection*, Arxiv Preprint (2013).
 - [Par01] Pablo A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*.
 - [Pie12] Krzysztof Pietrzak, *Cryptography from learning parity with noise*, Proceedings of the 38th International Conference on Current Trends in Theory and Practice of Computer Science (Berlin, Heidelberg), SOFSEM'12, Springer-Verlag, 2012, pp. 99–114.
 - [WBS14] Tengyao Wang, Quentin Berthet, and Richard J. Samworth, *Statistical and computational trade-offs in estimation of sparse principal components*, Preprint (2014).

APPENDIX A: PROOF OF THEOREM 6.2

PROOF. For $\Delta > \tilde{\Delta}_{k,\pi}$, taking $\alpha = \pi 2^{-(k+1)}$ in the results of Lemma 6.1 yields the desired upper bound, as $2\alpha^2\Delta - \log(2) > 0$.

For $\Delta < \tilde{\Delta}_{k,\pi}$, we derive a bound on the total variation distance $d_{\text{TV}}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted},\pi})$, through the inequality

$$d_{\text{TV}}(\mathbf{P}_{\text{unif}}, \mathbf{P}_{\text{planted},\pi}) = \frac{1}{2} \mathbf{E} \left[\left| \frac{\mathbf{P}_{\text{planted},\pi}(V) - 1}{\mathbf{P}_{\text{unif}}} \right| \right] \leq \frac{1}{2} \sqrt{\mathbf{E} \left[\left(\frac{\mathbf{P}_{\text{planted},\pi}(V) - 1}{\mathbf{P}_{\text{unif}}} \right)^2 \right]}.$$

The term inside the square root being equal to the chi-square divergence $\chi^2(\mathbf{P}_{\text{planted},\pi}, \mathbf{P}_{\text{unif}})$ between the two distributions. We write $\mathbf{P}_{x,\pi} = q_{x,\pi}^{\otimes m}$ and $\mathbf{P}_{\text{unif}} = q_0^{\otimes m}$ as products of the distribution of each independent V_j . Writing out $\mathbf{P}_{\text{planted},\pi}$ as a uniform mixture of the $\mathbf{P}_{x,\pi}$ yields

$$\begin{aligned} \chi^2(\mathbf{P}_{\text{planted},\pi}, \mathbf{P}_{\text{unif}}) &= \frac{1}{2^{2n}} \sum_{x, x' \in \mathbb{F}_2^n} \mathbf{E} \left[\frac{\mathbf{P}_{x,\pi}}{\mathbf{P}_{\text{unif}}} \frac{\mathbf{P}_{x',\pi}}{\mathbf{P}_{\text{unif}}} (V) \right] - 1 \\ &= \frac{1}{2^{2n}} \sum_{x, x' \in \mathbb{F}_2^n} \mathbf{E} \left[\frac{q_{x,\pi}}{q_0} \frac{q_{x',\pi}}{q_0} (V_1) \right]^m - 1 \\ &= \frac{1}{2^{2n}} \sum_{x \in \mathbb{F}_2^n} \mathbf{E} \left[\left(\frac{q_{x,\pi}}{q_0} (V_1) \right)^2 \right]^n + \frac{1}{2^{2n}} \sum_{x \neq x'} \mathbf{E} \left[\frac{q_{x,\pi}}{q_0} \frac{q_{x',\pi}}{q_0} (V_1) \right]^m - 1. \end{aligned}$$

Note that $q_{x,\pi} = (1 - \pi)q_0 + \pi q_x$, where q_x is the uniform distribution on k -flats that do not contain x (the planting distribution), so that

$$\frac{q_{x,\pi}}{q_0} = 1 + \pi \left[\frac{q_x}{q_0} - 1 \right].$$

Substituting this in the above yields

$$\chi^2(\mathbf{P}_{\text{planted},\pi}, \mathbf{P}_{\text{unif}}) = \frac{1}{2^{2n}} \sum_{x \in \mathbb{F}_2^n} \left(1 + \pi^2 \left[\mathbf{E} \left[\left(\frac{q_x}{q_0} (V_1) \right)^2 \right] - 1 \right] \right)^m + \frac{1}{2^{2n}} \sum_{x \neq x'} \left(1 + \pi^2 \left[\mathbf{E} \left[\frac{q_x}{q_0} \frac{q_{x'}}{q_0} (V_1) \right] - 1 \right] \right)^m - 1.$$

Furthermore, for any k -flat V_1 , it holds that $q_x/q_0(V_1) = (N/N_k)\mathbf{1}\{x \notin V_1\}$. We give the following upper bound the last two terms of this equation's RHS,

$$\begin{aligned} \frac{1}{2^{2n}} \sum_{x \neq x'} \left(1 + \pi^2 \left[\mathbf{E} \left[\frac{q_x}{q_0} \frac{q_{x'}}{q_0} (V_1) \right] - 1 \right] \right)^m - 1 &\leq \frac{1}{2^{2n}} 2^n \left(1 - \pi^2 + \pi^2 \frac{\mathbf{P}_{\text{unif}}(x, x' \notin V_1)}{(1 - 2^{-k})^2} \right)^m \\ &\leq \left(\frac{1 - \pi^2}{2} \right)^n \left(1 + \frac{\pi^2}{1 - \pi^2} \frac{2 + 2^{-k}}{1 - 2^{-k}} \frac{1}{2^n - 1} \right)^{\Delta n} - 1 \\ &\leq \left(1 + \frac{c_k \pi^2}{2^n - 1} \right)^{c_k n / \pi^2} - 1, \end{aligned}$$

for some constant $c_k > 0$ (independent of n and π), by the formula for $\mathbf{P}_{\text{unif}}(x, x' \notin V_1)$ derived in the proof of Lemma 3.2. The last term converges to 0 when $n \rightarrow +\infty$. We bound as well the first term of the main equation's RHS

$$\begin{aligned} \frac{1}{2^{2n}} \sum_{x \in \mathbb{F}_2^n} \left(1 + \pi^2 \left[\mathbf{E} \left[\left(\frac{q_x}{q_0} (V_1) \right)^2 \right] - 1 \right] \right)^m &\leq \frac{1}{2^{2n}} 2^n (1 + \pi^2 (\mathbf{P}_{\text{unif}}(x \notin V_1) - 1))^m \\ &\leq \frac{1}{2^n} \left(1 + \frac{\pi^2}{2^k - 1} \right)^{\Delta n}. \end{aligned}$$

Taking $\Delta < \Delta_{k,\pi} = 2^k \log(2)/\pi^2$ yields $1/2(1 + \pi^2/(2^k - 1))^{\Delta} < 1$, and all the terms of $\chi^2(\mathbf{P}_{\text{planted},\pi}, \mathbf{P}_{\text{unif}})$ go to 0 when $n \rightarrow +\infty$.

□

QUENTIN BERTHET
DEPARTMENT OF COMPUTING AND
MATHEMATICAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CA 99125, USA
(qberthet@caltech.edu)

JORDAN S. ELLENBERG
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WISCONSIN
MADISON, WI 53706, USA
(ellenber@math.wisc.edu)